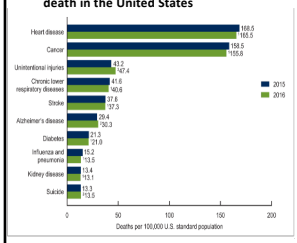
Complex Trait Association Analysis of Rare Variants Obtained from Sequence Data

© 2022 Suzanne M. Leal, suzannemleal@gmail.com

1

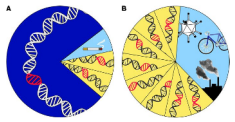
Complex Diseases (Traits)

Top 10 leading causes of death in the United States



D. Kenneth, et al. NCHS Data Brief No. 293, 2017

Genetic and environmental contribution to complex disorders



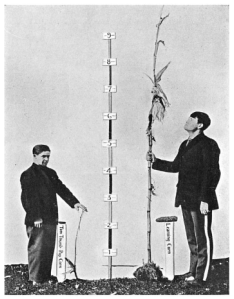
T.A. Manolio, et al. J clin Invest, 2001

2

Heritability for Common Traits

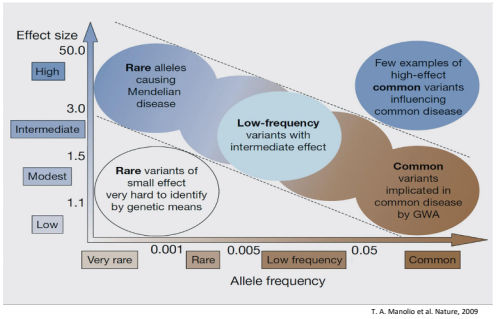
Human height heritability is ~80%

- Strongly associated common variation explain 21—29%
- All common variation explains 60% of height heritability



3

Allelic Architecture



T. A. Manolio et al. Nature, 2009

4

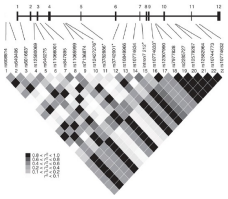
Complex Disease – Common Variant Associations

- Disease susceptibility is conferred by variants which are common within populations
  - Variants are old and widespread
- These variants have modest phenotypic effect
- This model is supported by a large number of replicated examples
  - Age Related Macular Degeneration (Klein et al. 2005)
    - Complement factor H (CFH) gene

5

Studying Complex Traits – Common Variant Associations

- Hundreds of thousands of Single nucleotide polymorphism (SNPs) genotyped and analyzed
  - Indirect mapping
    - Markers usually had a minor allele frequency (MAF) > 0.05
    - Usually not pathogenic – tag SNPs
    - In linkage disequilibrium with disease susceptibility variant



6

Complex Trait – Common Variant Associations

- Although highly successful in identifying thousands of complex trait loci
- Usually pathogenic susceptibility variant(s) not identified

7

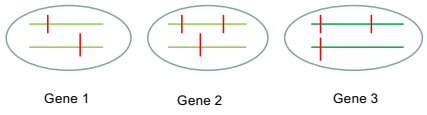
Complex Disease – Rare Variant Associations

- Complex traits are the result of multiple rare variants
  - Although first thought to large effects, there effect sizes are usually small
- Although these variants are rare, e.g., MAF<0.005
  - Collectively they may be quite common
- Direct tests of this hypothesis where first reported >15 years ago
  - Dallas Heart Study
    - Small sample ~1,200 individuals
      - Multi-ethnic
      - Used "extreme" sampling
    - Plasma low density lipoprotein levels (Cohen et al. 2004)
      - NPC1L1

8

Rationale for Rare Variant Aggregate Association Tests

- Testing individual variants with low effect sizes and minor allele frequencies (MAFs)
  - Underpowered to detect associations
- Testing variants in aggregate increases MAFs
  - Improving the power to detect associations



9

Caveats - Aggregate Rare Variant Association Tests

- Misclassification of variants can reduce power
  - Inclusion of non-causal variants
  - Exclusion of causal variants
- Analysis is limited to
  - Genes
  - Genes within pathways
- Analysis outside of exonic regions is problematic
  - Unlikely a sliding window approach will work
    - Size of window unknown and will differ across the genome
  - A better understanding of functionality outside the coding regions is necessary
    - Predicted functional regions, enhancer regions, transcription factors, DNase I hypersensitivity sites, etc.

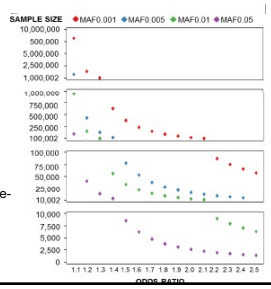
10

Analysis of Rare Variants

- For biobank sized datasets higher frequency rare variants, e.g., 0.5% can be analyzed individually
  - Using same same methods implemented for common variants

Example
$\alpha=5 \times 10^{-8*}$
Disease prevalence 5%
$1-\beta = 0.80$

*Note: a more stringent significance criterion may be necessary for genome-wide sequence data. Due to a larger number of effective tests compared to analysis of common variant GWAS panels



11

A Few Rare Variant Association Tests

- Combined Multivariate Collapsing (CMC)
  - Li and Leal AJHG 2008
- Burden of Rare Variants (BRV)
  - Auer, Wang, Leal Genet Epidemiol 2013
- Weighted Sum Statistic (WSS)
  - Madsen and Browning PloS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Liu and Leal PloS Genet 2010
- Variable Threshold (VT)
  - Price et al. AJHG 2010

Fixed Effect Tests

- Sequence Kernel Association Test (SKAT)
  - Wu et al. AJHG 2011

Random Effect Test

- SKAT-O
  - Lee et al. AJHG 2012

Optimal test

12

Types of Aggregate Analyses

- Frequency cut offs used to determine which variants to include in the analysis
  - Rare Variants (e.g., MAF<0.05% frequency)
  - Rare and low (MAF=0.05-5%) frequency variants
- Maximization approaches
- Tests developed to detection associations when variants effects are bidirectional
  - e.g., protective and detrimental
- Incorporate weights based upon annotation
  - Frequency
    - e.g., gnomAD
  - Functionality
    - CADD c-scores

13

Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Combined multivariate & collapsing (CMC)
  - Li & Leal, AJHG 2008
- Collapsing scheme which can be used in the regression framework
  - Can use various criteria to determine which variants to collapse into subgroups
    - Variant frequency
    - Predicted functionality

14

CMC

- Define covariate Xj for individual j as

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

- Compute Fisher exact test for 2x2 table

Number of cases for which one or more rare variants are observed e.g., nonsynonymous variants freq. ≤1%

	X=1	X=0
cases		
controls		

Number of cases without a rare variants

Number of controls for which one or more rare variants are observed

Number of controls without a rare variants

Can also use same coding in a regression framework

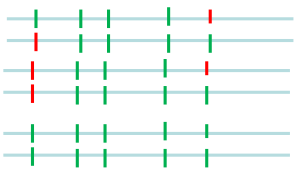
15

CMC

- Example of coding used in regression framework:
  - Binary coding

$$X_j = \begin{cases} 1 & \text{if rare variants present} \\ 0 & \text{otherwise} \end{cases}$$

  - Gene region with 5 variant sites
  -



Individual	Coding
1	1
2	1
3	0

Rare Variant Sites

Green bars: Major allele is observed in the study subject

Red bars: Minor allele has been observed

16

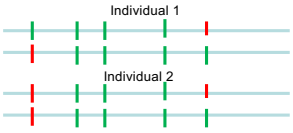
Methods to Detect Rare Variant Associations Using Variant Frequency Cut-offs

- Gene-or Region-based Analysis of Variants of Intermediate and Low frequency (GRANVIL)
  - Aggregate number of rare variants used as regressors in a linear regression model
  - Can be extended to case-control studies
    - Morris & Zeggini 2010 Genet. Epidemiol
  - Test also referred to as MZ

17

GRANVIL

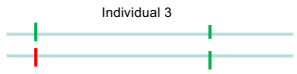
- Example of coding used in regression framework
  - Gene region with 5 variant sites – data available on all sites
  -



Coded 2/5 (0.4)

Coded 2/5 (0.4) Note same coding for heterozygous and homozygous genotypes

- Missing data for three of the five variant sites



Coded 1/2 (0.5)

Burden Rare Variant (BRV) extension (Auer et al. 2013 Genet Epidemiol)

Individual 1: Coded 2

Individual 2: Coded 3

Individual 3: Coded 1

18

Methods to Detect Rare Variant Associations Weighted Approaches

- Group-wise association test for rare variants using the Weighted Sum Statistic (WSS)
  - Variants are weighted inversely by their frequency in controls (rare variants are up-weighted)
    - Madsen & Browning, PLoS Genet 2009
- Kernel based adaptive cluster (KBAC)
  - Adaptive weighting based on multilocus genotype
    - Liu & Leal, PLoS Genet 2010

19

Methods to Detect Rare Variant Associations Maximization Approaches

- Variable Threshold (VT) method
  - Uses variable allele frequency thresholds and maximizes the test statistic
  - Can also incorporate weighting based on functional information
    - Price et al. AJHG 2010
- RareCover
  - Maximizes the test statistic over all variants with a region using a greedy heuristic algorithm
    - Bhatia et al. 2010 PLoS Computational Biology

20

Methods to Detect Associations with Protective & Detrimental Variants within a Region

- C-alpha
  - Detects variants counts in cases and controls that deviate from the expected binomial distribution
    - For qualitative traits only
      - Neale et al. 2011 PLoS Genet
- Sequence Kernel Association Test (SKAT)
  - Variance components score test performed in a regression framework
    - Can also incorporate weighting
  - Wu et al. 2011 AJHG

21

Optimal Test

- SKAT-O
  - Maximizes power by adaptively using the data to combine a burden test and the sequence kernel association tests
    - Lee et al. 2012 AJHG

22

Significance Level for Rare Variant Association Tests

- For exome data where individual genes are analyzed usually a Bonferroni correction for the number of genes tested is used
  - There is very little to no linkage disequilibrium between genes
- Bonferroni correction used
  - e.g., $p < 2.5 \times 10^{-6}$ (Correction for testing 20,000 genes)

23

Determine MAF Cut-offs for Aggregate Rare Variant Association Tests

- MAF cut-offs are frequently used to determine which variants to analyze in aggregate rare variant association tests
- MAF from controls should not be used
  - Increases in type I error rates
- Determine variant frequency cut-offs from databases
  - gnomAD
    - http://gnomad.broadinstitute.org/

24

Problem of Missing Genotypes for Aggregate Rare Variant Association Tests

- Same frequency of missing variant calls in cases and controls
  - Decrease in power
- More variant calls missing for either cases or controls
  - Increase in Type I error
  - Decrease in power
- Remove variant sites which are missing genotypes, e.g., >10%
- Can impute missing genotypes using observed allele frequencies
  - For the entire sample
    - Not based on case or control status
- Analyze imputed data using dosages

25

Dosages

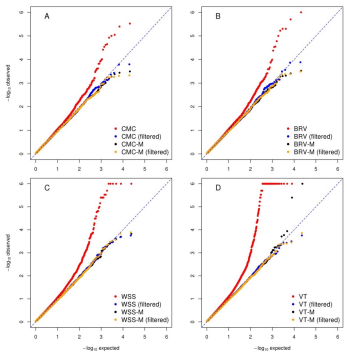
- Genotypes are no longer assigned 0 (1/1), 1 (1/2) or 2 (2/2)
  - Due to uncertainty
- Each genotype is assigned a probability
  - Probabilities sum to 1
- For example
  - Probability of 0 (1/1) genotype is 0.98 and 1 (1/2) genotype is 0.015
- The dosage can be estimated for this example as follows

0 x 0.98 = 0
1 x 0.015 = 0.015
2 x 0.005 = 0.01
Dosage = 0.025

- Instead of using the most likely genotype the dosage is used

26

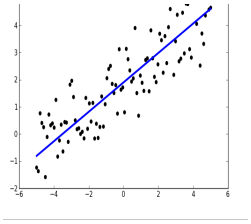
Results



27

Rare Variant Aggregate Methods

- Ideally should be performed in a regression framework
  - Logistic
  - Linear regression

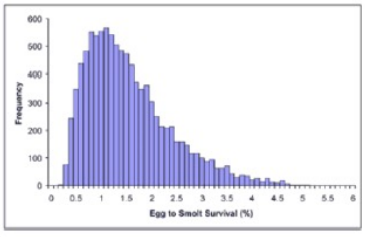


- Almost all methods have been extended to be implemented within a regression framework

28

Analyzing Quantitative Variants

- Most rare variant aggregate analysis methods can be performed on quantitative traits
- If phenotype data includes outliers or deviates from normality
  - Can increase type I errors



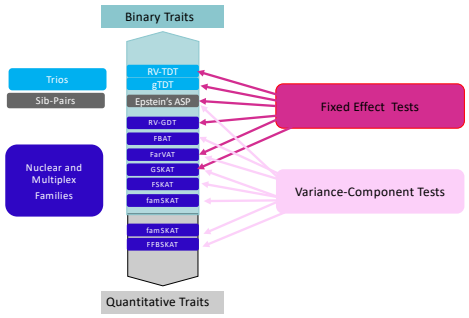
29

Analyzing Quantitative Variants

- For data that deviates from normality
  - Quantile-quantile normalization
- For data that includes outliers
  - Winsorize
- Don't winsorize and then normalize
- Instead of analyzing quantitative trait values
- Residual can be generated
  - If there are confounders which need to be controlled
    - Residuals are generated where confounders have been adjusted

30

Family-based Methods for Rare Variant Aggregate Association Analysis



31

Linear Mixed Model (LMM) & generalized LMM (GLMM) Analysis of Related & Unrelated Individuals

- LMM is an extension of a linear model to allow for both fixed & random effects and allows for non-independence of samples
  - Early implementations calculated the kinship matrix Φ on the basis of known relationships
  - Amin et al. (2007) proposed to estimate kinships based on genome-wide variant data
    - The generalized relationship matrix (GRM) can be estimated for all individuals using for example identical-by-descent (IBD) sharing
- Extended to binary (case-control) traits - GLMM

32

LMM and GLMM: Analysis of Related & Unrelated Individuals

- Can be applied to analyze families, cryptically related, & unrelated individuals
  - e.g., UK Biobank
    - 500K study subjects of which 30.3% are ≤ 3rd degree relatives & 4.5% sib-pairs
- More recent implementation for large scale data using a variety of methods
  - BOLT-LMM (Loh et al. 2015)
  - FastGWA (Jiang et al. 2019)
  - SAIGE (Zhao et al. 2015)*
  - REGENIE (Mbatchou et al. 2020) *
  - SMMAT (Chen et al. 2019)
- *Can be used to analyze data where case to control ratio is very unbalanced
  - e.g., 20 cases for every control

33

LMM and GLMM: Analysis of Related & Unrelated Individuals

- To allow for use with very biobank sized data
- REGENIE does not use the GRM
  - It uses whole genome regression instead, i.e., the ridge regression
    - In essence, it includes all the SNVs as covariates in the null model
      - Performed by blocks to avoid having to load the entire genome in memory
        - » Using different effect size differences per block
- This large-scale approximation may not control type I error for individuals that are closely related
  - e.g., when only families are being analyzed
- Can use for example SMMAT
  - Which uses the GRM

34

LMM and GLMM: Analysis of Related & Unrelated Individuals

- A few programs which can perform rare variant aggregate analysis
  - REGENIE - Burden test
  - SMMAT - Burden, SKAT, & SKAT-O tests
  - rvtests (Zhan 2020) implements BOLT-LMM to perform burden association analysis
- An alternative for rare variant aggregate analysis
  - Recode variants within gene regions and then analyze

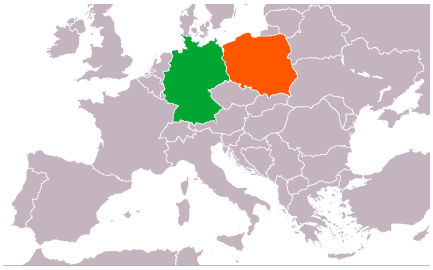
35

Rare Variant Aggregate Methods

- Can control for covariates in the analysis which are potential confounders
  - Age
  - Sex
  - Body Mass Index (BMI)
  - Smoking pack years

36

Confounder -Population Substructure and Admixture



37

Population Substructure and Admixture

- If proportion of cases and controls sampled from each population is different
  - Can occur due to
    - Disease frequency is different between populations
    - Sloppy sampling
- Population substructure\admixture can cause detection of differences in variant frequencies within a gene which is due to sampling and not disease status
  - False positive findings can be increased

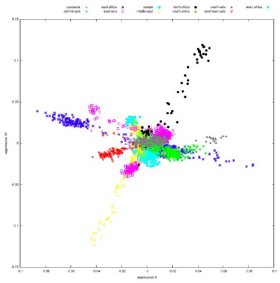
38

Example Pima Indians

39

Population Substructure and Admixture

- Currently PCA or MDS are use to control for population substructure\admixture
  - Controls on the global level
  - May not be sufficient in particular for admixed populations



40

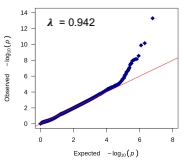
Rare Variant Aggregate Association Analysis

- When analyzing different populations, e.g.,
  - Africans
  - Europeans
- Always analyze each group separately
- Meta-analysis can be used to combine the results from each group
- May also want to use meta-analysis to combine data that was generated differently
  - e.g., to combine two or more batches

41

Rare Variant Aggregate Methods

- Best to obtain components to include in the regression model (including LMM and GLMM)
  - using variants which are not in LD e.g., r^2<0.5 (pruned)
  - covering a wide range of the allelic frequency spectrum e.g., >0.1%
  - Evaluate how many components need to be included
    - Don't include a fix number of components
      - e.g., 5 or 10 components
- Success of PCA\MDS in controlling for population substructure\admixture can be evaluated through lambda and examining Quantile-Quantile (QQ) plots



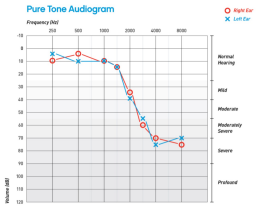
42

Part IV
Analysis of UK Biobank Exome Data to Study the Etiology of Late-onset Hearing Impairment

43

Age-related Hearing Impairment (ARHI) (aka Presbycusis)

- ARHI can impact quality of life and daily functioning
- ARHI is one of the most common adult conditions
  - In the USA
    - ARHI affects 50% of individuals >75 years of age
    - It is estimated that 30-40 million will be affected with significant ARHI by 2030



44

UK Biobank

- 500,000 individuals randomly sampled
  - Aged 40-69 at time of enrollment
    - To be followed for 20 years
    - Predominantly Europeans
      - Also includes South Asians and individuals of African Ancestry and smaller number of individuals of a few other ancestries
- Extensive phenotype data*
  - Qualitative and quantitative traits
    - ICD-10 and ICD-9 codes
    - Self reports
    - Cognitive test
    - Brain MRIs
    - NMR-metabolomics data

*Data showcase can be used to examine phenotypes and sample sizes available

45

UK Biobank – Genetic Data

- Genotype arrays
  - 500K
- Imputed variants using the Haplotype Reference Consortium (HRC) panel
  - 500K
- Imputed variants using TOPMed reference panel
  - to be released
- Exome sequence data
  - Currently 300K
  - 450K in 2022
- Whole genome sequence data
  - to be released
- Telomere length
  - to be released

46

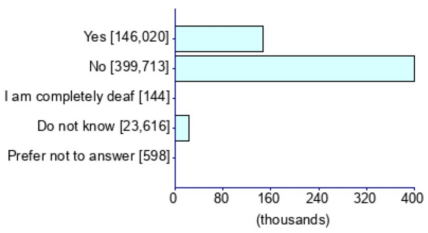
UK Biobank

- Publicly available
  - Submit application stating purpose of the study to request data
- Our current studies
  - Age related hearing impairment and tinnitus
    - Main effects
    - Interactions
    - Pleiotropy
  - Pleiotropy
    - Asthma, adiposity, lipids, hypertension
  - Alzheimer disease
    - Use of surrogate phenotypes
      - e.g., family history of AD

47

Hearing difficulty/problems -Data field 2247

569,977* items of data are available, covering **498,704** participants



*Due to repeat visits

48

Repeat measures*

- Individuals with inconsistent answers removed

	Visit 1	Visit 2	Visit 3	Visit 4	
Study subject A	Problems Hearing	No Hearing Problems	No Hearing Problems	No Hearing Problems	Inconsistent Remove
Study subject B	No Hearing Problems	No Hearing Problems	Problems Hearing	Problems Hearing	Consistent (Case)
Study subject C	No Hearing Problems	No Hearing Problems	No Hearing Problems	No Hearing Problems	Consistent (Control)

Cases – age of onset is the first visit where hearing problem was reported
Controls – age at last visit

*Majority of study subjects currently have data from only one visit

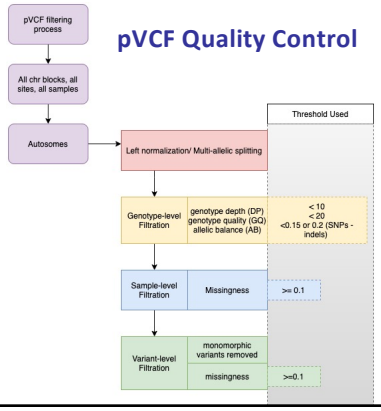
49

Exclusion Criteria
Obtained from ICD10, ICD9, & Self Report

- Deafness
- Early-onset hearing impairment
- Otosclerosis
- Meniere's
- Labyrinthitis
- Disorders of acoustic nerve
- Bell's palsy
- History of chronic suppurative and nonsuppurative otitis media
- Meningitis
- Encephalitis, myelitis, and encephalomyelitis
- Etc.

50

pVCF Quality Control



51

Filtering Strategy

N=200,619
Individuals with exomes and phenotype data

N=200,386
Individuals consistent for sex

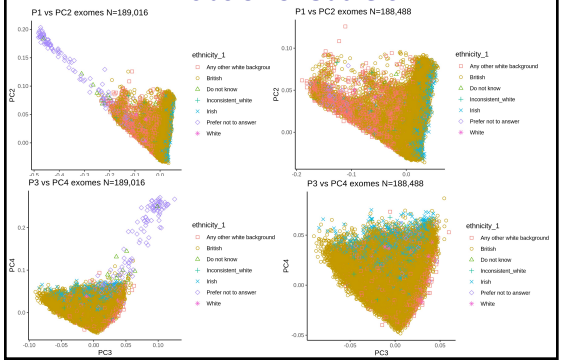
N=189,001
Individuals self-report to be of European ancestry

N=187,908
After removal of outliers based on PCA analysis – Mahalanobis distance (0.997)

N=180,318
After applying phenotype exclusion criteria (ICD10, ICD9, & self-report)

52

Principal Components Analysis and Exclusion of Outliers



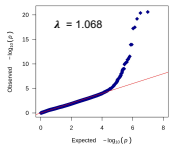
53

Analysis of Exome Data

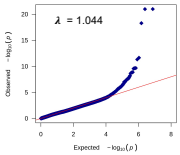
- Analysis performed using GLMM (REGENIE)
  - To control for inclusion of related individuals
  - Genotype array data (~800,000 variants) were used for the ridge regression
    - Data pruned to remove variants with a r^2>0.1
      - Using exome data for the ridge regression led to an an inflated lambda value

Hearing Difficulty Data Field 2247

QQ Plot using exome data for ridge regression



QQ Plot using genotype data for ridge regression



54

Analysis of Exome Data – Single Variant

- Sex, age, and two PCAs included as covariates
  - Age for cases first report of hearing problem and controls age at last visit
  - The PCAs where recalculated for only individuals included in the analysis
    - Using the pruned genotypes array data (r^2>0.1)
    - PCAs were calculated using FLASHPCA2 (Abraham et al. 2017)
- All exome variants with at least four alternative alleles were analyzed

55

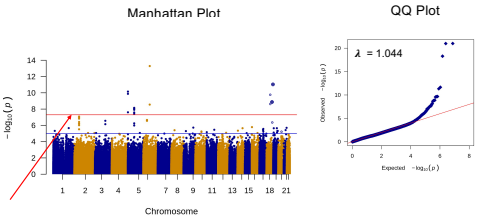
Analysis of Exome data – Single Variant

- Sample was split
  - First release of 50K exome were uses as the replication* sample
  - Second release of 150K exome were used as the discovery sample
- Entire exome sample (200K) analyzed
  - Replication* sample genotype and imputed data (HRC) for 300K individuals who were not included in the exome data
  - For the imputed data all variants with an INFO score > 0.3 were analyzed

*For the replication it is not necessary to use a genome-wide significance level of 5 x 10^{-8} for single variant tests or 2.5 x 10^{-6} for gene-based rare variant aggregate analysis. Significance level is adjusted for the number of variants/genes tested in the replication sample

56

Hearing Difficulty - Data Field 2247

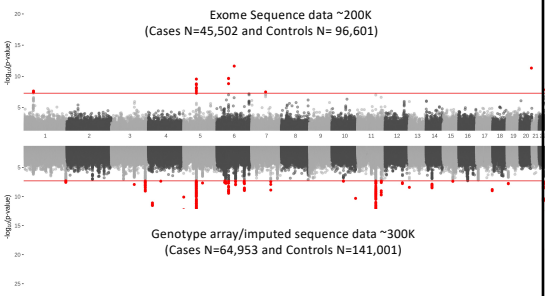


Genome-wide significance level 5 x 10^{-8} (red line)

Cases N=45,502
Controls N= 96,601

57

Hudson Plot Discovery and Replication Hearing Difficulty Data Field 2247



58

REGENIE Rare Variant Aggregate Analysis

- Rare variants can be analyzed in aggregate within a gene or region
- Three different codes can be used
  - Max
  - Sum
  - Comphet
    - Although called Comphet this term is not correct because the phase is unknown
      - Variants may be on the same haplotype

Single variant sites	max	sum	comphet
00000000000000 →	0	0	0
00000100010000 →	1	2	2
00201011010100 →	2	7	2

https://rgcgithub.github.io/regenie/options/

59

Selection of Variants to Include in Rare Variant Aggregate Association Tests

Annotation File		Mask File		AAF file
1:55039839:T:C PCSK9 LoF 1:55039842:G:A PCSK9 missense	+	Mask1 LoF Mask2 LoF,missense	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06
1:55039839:T:C PCSK9 CADD30 1:55039842:G:A PCSK9 CADD20	+	Mask1 CADD score > 30 Mask2 CADD score > 20	+	1:55039839:T:C 1.53e-05 1:55039842:G:A 2.19e-06

REGENIE will use information from the annotation and alternative allele frequency (AAF) files to build the Masks (variants to be included in the association testing) and also use the select coding (max, sum, or comphet) to perform the rare variant aggregate association testing

60

Analysis of Exome Data
Rare Variant Aggregate Analysis

- Analysis performed using REGENIE
  - To control for inclusion of related individuals
  - Genotype array data (~800,000 variants) were used to construct for the ridge regression
    - Data pruned to remove variants with a r^2>0.1
- Sex, age and two PCAs included as covariates
  - The PCAs where recalculated for only individuals included in the analysis
    - Using the pruned genotypes array data (r^2>0.1)

61

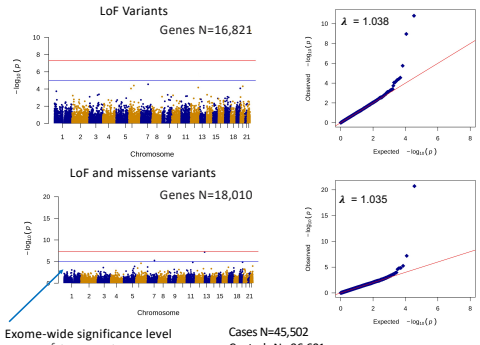
Analysis of Exome Data
Rare Variant Aggregate Analysis

- Genes with at least two variants were analyzed
- Max coding was used
- Two masks were used
  - Mask 1 - LoF variants
  - Mask 2 – LoF and missense variants
- Two minor allele frequency cut-offs were used
  - The frequencies for each variant site were obtained from gnomAD non-Finnish Europeans
  - <0.005
  - <0.01*
- Sample was split
  - First release of 50K exome were uses as the replication sample
  - Second release of 150K exome were used as the discovery sample
- Entire exome sample (200K) was analyzed*

*Results shown for this scenario

62

Hearing Difficulty - Data Field 2247

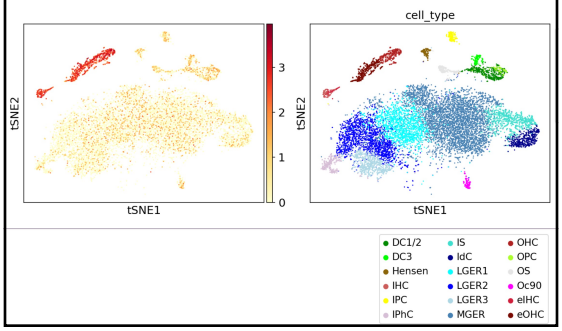


Exome-wide significance level
2.5 x 10^{-6} (blue line)

Cases N=45,502
Controls N= 96,601

63

Expression - gEAR



64